基于可拓小生境量子粒子群算法的特征选择*

李志鹏　李卫忠
(空军工程大学防空反导学院　西安 710051)

摘要: 【目的】对适用于特征选择的算法进行研究, 有效提高文本分类精度和效率。【方法】结合特征选择特点, 以可拓理论为基础构造小生境量子粒子群算法, 通过改进增强算法搜索能力, 将不同的特征选择方法用于文本分类并进行比较。【结果】实验结果表明, 与 IG、MI 等方法相比, 基于可拓小生境量子粒子群算法的特征选择在文本分类中取得了较好效果, 算法的求解精度得到明显提升。【局限】所提出的特征选择方法在时间效率上有待改善。【结论】对量子粒子群算法的改进措施有效提高了算法的搜索能力, 在特征选择的应用中达到较好的效果。

关键词: 特征选择　量子粒子群　可拓理论　小生境　适应度共享
分类号: TP301

1 引　言

随着智能信息技术的发展, 海量高维数据处理成为机器学习领域面临的极大挑战之一, 数据维度的增加导致解空间规模呈指数级增长, 加之现实数据中包含着大量冗余信息特征, 传统的机器学习算法已无法满足高维、稀疏数据的处理要求[1]。为提高数据处理的效率, 国内外对特征提取与选择技术进行了深入研究, 目前大多数特征选择方法采用启发式搜索, 在牺牲搜索空间的基础上提高了搜索效率, 但精度也因此受到一定影响。粒子群算法实现简单, 不少文献对其在特征选择问题求解中的应用进行了研究, 但研究重点多集中于算法自身的搜索能力, 对特征选择问题的性质、特点及算法的易懂性、可操作性考虑较少[2-4]。本文结合特征选择特点, 提出一种应用小生境和反向学习策略的改进量子粒子群算法(Quantum-behaved Particle Swarm Optimization Algorithm Using Niche and Opposition-Based Learning, NOL-QPSO), 以可拓理论为基础改进粒子群算法模型, 结合粗糙集理论、适应度动态共享技术, 引入精英反向学习策略, 有效解决算法过早收敛的问题, 增强寻优能力, 以更好地实现文本特征子集的选取。

2 研究现状

特征选择的目的是从某组特征集中选择出若干个最具代表性的有效特征组成具有类别区分能力的特征子集, 对样本进行识别、分类, 从而降低特征空间的维数。随着信息领域数据规模的增大和特征维数的增加, 传统的学习算法往往会表现出性能上的局限性, 因此, 国内外学者近年来对特征选择技术和方法进行了广泛而深入的研究。在文本分类中, 常用的有粗糙集、支持向量机、决策树、神经网络及基于蚁群、粒子群等群智能算法的分类方法。文献[5]通过引入遗传算法, 对传统的特征提取方法进行改进, 在用于特征维数较少的情况时取得较好的效果。文献[6]提出的基于野草算法的文本特征选择方法, 通过平衡词条权重与选择机率, 增强了特征选择结果的准确性。目前, 粒子群算法在文本特征选择中的应用可以大致分为三个类别。

通讯作者: 李志鹏, ORCID: 0000-0002-7814-8924, E-mail: lizhipeng0888@yeah.net。

*本文系国家自然科学基金项目"基于 ELM 和 D-S 证据理论的'低慢小'目标识别中的不确定信息融合方法研究"(项目编号: 61503407)的研究成果之一。

(1) 通过粒子群算法对分类算法进行优化。文献[7]将粒子群算法与 K 近邻法结合，并应用于文本分类，在保证精度的基础上提高了复杂文本的分类速度。文献[8]通过粒子群优化算法较强的随机搜索能力对样本中的K近邻搜索，有效避免了粒子速度的影响，实验表明该方法与 KNN 算法相比分类精度更高。文献[9]提出一种基于PSO优化支持向量机的方法，通过改进的 PSO 算法对 SVM 参数进行优化，经过样本训练得到分类器，从而实现文本的分类。

(2) 将粒子群算法作为分类算法建立分类器，实现文本的分类。文献[10]结合频率统计函数、适应度函数和打分函数，通过粒子群算法确定出具体分类规则，实现了教育管理系统中文本资源的自动分类。文献[11]将混沌二进制粒子群算法与KNN算法结合，通过粒子群算法进行特征选择，并在此基础上利用KNN算法完成文本分类，分类准确率、召回率都有所改善。文献[12]对微粒群算法进行改进，提出一种混沌微粒群算法，并用于分类规则的提取。

(3) 直接利用粒子群算法完成特征选择。为压缩文本挖掘所占用的内存空间，提高算法速度，文献[13-14]提出了一种结合并行算法和二进制免疫量子粒子群算法的特征选择方法，不仅能准确获取特征子集，而且提高了算法的时间效率。文献[15]对异质数据的特征选择问题进行研究，提出一种基于多目标微粒群优化的特征选择方法，通过典型数据集的实验验证了方法的有效性。

粒子群算法作为一种简单、实用的优化算法，对特征选择问题的研究具有重要作用，进一步的研究不仅要从粒子群优化算法本身的改进入手，还要对算法应用于特征选择问题的途径和方式进行分析。

3　基于可拓理论的小生境构造策略

针对粒子群算法容易陷入局部最优和出现早熟等缺点，本文考虑通过引入小生境策略对粒子群优化算法的寻优能力进行优化，其基本思想是根据某种规则将种群划分为若干类，将解空间分为不同的搜索域，对不同的局部最优点展开同步搜索，避免过早收敛或过度搜索现象[16]。该技术首先要考虑的关键问题是小生境的划分，常用的方法是通过设定小生境半径划分子空间，这种方法虽简单可行，但对于复杂问题的求解效果并不理想，为此本文以物元可拓理论[17]为基础，采用可拓聚类算法构造小生境。

设 $Q(k)=\{X_i, i=1,2,\cdots,N\}$ 为含有 N 个样本的初始种群，每个样本特征维数为 n，样本 I 的数据模型可表示为:

$$X_i=[I,C,V]=\begin{bmatrix} I & C_1 & x_i^1 \\ & C_2 & x_i^2 \\ & \vdots & \vdots \\ & C_n & x_i^n \end{bmatrix}$$

令 $\min_j=\min\limits_{i=1}^{N} x_i^j$，$\max_j=\max\limits_{i=1}^{N} x_i^j$，则特征 C_j 的可行域可表示为: $V_j=[\min_j,\max_j]$。

定义 1　根据文献[18]，X_i 对类 S_l 的关联度 $K_x(S_l)$ 计算准则定义为:

$$\begin{cases} K_x(S_l)=\sum\limits_{j=1}^{n}\lambda_l^j k_l(x_i^j) \\ k_l(x_i^j)=\begin{cases} \dfrac{x_i^j-\min_l^j}{M_l^j-\min_l^j}, x_i^j\leqslant M_l^j \\ \dfrac{\max_l^j-x_i^j}{\max_l^j-M_l^j}, x_i^j>M_l^j \end{cases} \\ M_l^j=\dfrac{1}{m_l}\sum\limits_{i=1}^{m_l}x_i^j \end{cases}$$

其中，$\lambda_l=[\lambda_l^i]=[\lambda_l^1,\lambda_l^2,\cdots,\lambda_l^n]$ 为 n 项特征指标的权重，反映了各特征在个体评价中的重要程度。

为对各类的相似度作出评价，参照样本与类之间关联度的定义，本文对不同类之间的关联度(即类间关联度)及类与自身中心的关联度(即自关联度)作出定义:

定义 2　取类 S_Γ 的中心物元: $R_o=\begin{bmatrix} I & C_1 & M_\Gamma^1 \\ & C_2 & M_\Gamma^2 \\ & \vdots & \vdots \\ & C_n & M_\Gamma^n \end{bmatrix}$，类 S_Γ 与 $S_l,(l=1,2,\cdots,k且l\neq\Gamma)$ 的类间关联度可以通过 R_o 对类 S_l 的关联度 $K_o(S_l)$ 表示。$K_o(S_l)$ 值越大表明类 S_Γ 与 S_l 的相似性越大。

定义 3　当 $l=\Gamma$ 时，R_o 与类 S_Γ 的关联度则为类 S_Γ 的自关联度，表示为:

$$\overline{K}_{\Gamma}=\frac{1}{n_{\Gamma}}\sum_{i=1}^{n_{\Gamma}}K_i(S_{\Gamma})$$

其中，n_{Γ} 是类 S_{Γ} 的样本数目，$K_i(S_{\Gamma})$ 是类 S_{Γ} 第 i 个样本的关联度。

小生境具体构造方法如下：

①随机选取 k 个待聚类样本作为中心，形成 k 个初始类 $S_l,(l=1,2,\cdots,k)$。

②采用文献[19]的方法对其余样本进行聚类，通过样本关联度表征样本 X_i 与类 S_l 的关联程度，以此判定其具体类属：令 $K_x(S_r)=\max\limits_{i=1}^{k}K_x(S_i)$，若 $K_x(S_r)\geqslant 0$，则 $X_i\in S_r$；若 $K_x(S_r)<0$，则将 X_i 划为新类 S_{Γ}，$k=k+1$。

③为使类的划分得到约简，考虑将相似性较强的类归并，根据类间关联度，若 $K_o(S_l)>\overline{K}_{\Gamma}$，则将类 S_{Γ} 并入 S_l，$k=k-1$。

④聚类调整完成后，更新各类中心物元，重新计算关联度，重复新一轮的样本归类，直至归类结果不变。

通过以上基于可拓理论的聚类过程，形成稳定多样的小生境，各个子种群在相对独立的空间寻优，可以避免粒子群陷入局部极值，增强算法的全局性。

4 小生境量子粒子群优化的特征选择

4.1 个体编码

在特征选择中，个体编码通常采用二进制编码的方式，一个0-1字符串代表一个粒子，每一数位对应一种特征，其中 0 表示该数位对应的特征在粒子中不被选择，而 1 数位对应的特征包含在粒子中。原始特征集通过编码转化为由长度为 n 的 0-1 字符串组成的解空间。如四维样本 0-1-1-0 表示个体选择了特征 C_2 和 C_3，而摒弃特征 C_1、C_4。

4.2 粒子更新

在量子粒子群算法中，粒子被赋予量子行为，其状态通过波函数来描述。粒子在量子 δ 势阱的基础上不断向局部吸引点靠近，粒子出现在空间某一处的概率通过求解薛定谔方程得出，从而利用蒙特卡罗模拟更新粒子位置，其方程具体描述为[13]：

$$\begin{cases}x_i(g+1)=\tilde{x}(g)\pm\dfrac{L}{2}\ln(\dfrac{1}{u})\\ \tilde{x}(g)=\varphi x_{i-best}(g)+(1-\varphi)x_{best}(g)\\ L=2\beta\left|x_i(g)-\overline{x(g)}\right|\\ \overline{x(g)}=\dfrac{1}{N}\sum\limits_{i=1}^{N}x_i(g)\end{cases}$$

其中，N 表示粒子的个数；L 为势阱长度；$u,\varphi\in[0,1]$ 为随机数；$x_{i-best}(g)$ 是粒子 i 经过 g 次迭代的历史最优位置；$x_{best}(g)$ 是粒子群经过 g 次迭代的全局最优位置；$\tilde{x}(g)$ 为局部吸引点，在 $x_{i-best}(g)$ 和 $x_{best}(g)$ 间随机取得；$\overline{x(g)}$ 为粒子群的中心位置，即第 g 次迭代时粒子群个体的平均最佳位置；β 为收缩扩张系数，在实际应用中往往需要动态赋值，其变化取值可定义为：

$$\beta=\begin{cases}2\cdot\dfrac{f_{best}}{f(i)},&\dfrac{f_{best}}{f(i)}<0.5\\ \dfrac{f_{best}}{f(i)}+1,&\dfrac{f_{best}}{f(i)}\geqslant 0.5\end{cases}$$

其中，$f(i)$ 表示个体 i 的适应度，f_{best} 表示当前最优个体的适应度值。

4.3 适应度动态共享

在特征选择中，选取的属性个数越少，则属性对决策产生的影响即支持度越高，应为其赋予更大的适应度值，本文据此结合粗糙集理论对适应度函数定义如下：

$$f(i)=\begin{cases}\dfrac{Card(C-B(i))}{Card(C)}\gamma_B(Q),&\gamma_C(Q)-\gamma_B(Q)\geqslant\varepsilon\\ \dfrac{2\cdot Card(C-B(i))}{Card(C)}\gamma_B(Q),&\gamma_C(Q)-\gamma_B(Q)<\varepsilon\end{cases}$$

其中，ε 为误差参数；$B(i)$ 表示个体 i 中对应位为 1 的属性组成的属性集。当 $\gamma_C(Q)-\gamma_B(Q)<\varepsilon$ 时，$B(i)$ 接近最优解，此时适度赋予个体 i 更优的适应度，以获取较好的优化结果。适应度函数涉及粗糙集基础理论，对必要的概念作出如下简要说明，详见文献[20]。

定义 4 设信息系统 $S=\langle U,C\cup Q,V,f\rangle$，若有 $U/C=\{X_1,X_2,\cdots,X_n\}$，$U/Q=\{Y_1,Y_2,\cdots,Y_m\}$，则 Q 对 C 的支持度为：$\gamma_C(Q)=\dfrac{1}{|U|}\sum\limits_{i=1}^{m}|Pos_C(Y_i)|$，$Y_i\in U/Q$；若有 $B\subseteq C$，$\gamma_B(Q)=\gamma_C(Q)$，则 B 为 C 的约简集。将属性 a 纳入 $R\subseteq C$，对 U/Q 的重要度为：$SGF(a,R,Q)=\gamma_{R+\{a\}}(Q)+\gamma_R(Q)$。重要度的大小反映了 a 在已知条件 R 下对决策 Q 的影响程度。

在粒子群算法优化过程中，距最优点越近的粒子，其位置更新会受到越大限制，导致粒子群只能在局部极值邻域进行搜索，这是使粒子陷入局部最优和影响算法寻优精度的主要因素之一。因此，本文考虑在算

法中引入共享函数[21-22]，当搜索陷入局部最优时，以汉明距离为依据，选取距局部最优解较近的粒子，通过调整其适应度，促使粒子尽快逸出早熟区。但在迭代过程中，适应度更新后的个体可能会再次陷入原早熟区，为有效解决这一问题，本文利用共享距离 D 划定共享区，利用适应度动态共享策略对共享区内个体的适应度进行调节。

设某次迭代中，含有若干粒子的群体收敛于 $X_{local-best}$，个体 i 到 $X_{local-best}$ 的距离用 Block 距离 d_i 表示：

$$d_i = \sum_{j=1}^{n} \left| x_i^j - x_{local-best}^j \right|$$

对个体 i，若 $d_i < D$，则说明其进入共享区，将其初始化为新粒子，并将适应度更新为共享适应度 $f(i)_{new}$。本文对 $f(i)_{new}$ 定义如下：

$$f(i)_{new} = H \cdot \frac{D}{d_{new}} f(i), \quad H\text{为常数}$$

其中，d_{new} 为新粒子距局部最优解 $X_{local-best}$ 的距离；H 为修正权值，可根据实际情况调整值的大小。不难看出，d_{new} 越小，粒子距局部最优解越近，为其赋予越大的适应度，能够使个体以更大概率突破局部限制，避免新群体再次陷入局部最优。

4.4 精英反向学习策略

在量子粒子群算法中，全局最优粒子往往会包含更多引导种群向全局最优收敛的价值信息，对种群的进化方向具有重要的引导作用。最优粒子在当前种群中的自我学习能力受限，会影响算法的全局搜索能力，为拓展到当前种群以外的搜索空间对全局最优粒子进行深度挖掘，本文引入精英反向学习策略，粒子群每达到一定迭代次数，就对全局最优粒子进行一次反向学习，增强解空间的开发，提高算法精度。

反向学习[23]是通过求解当前解的反向解，并从当前解与其反向解中选取较优的作为新粒子参与下一代优化。设小生境 Q 中某个体 X_i 的反向粒子为

$$\overleftarrow{X_i} = \begin{bmatrix} \vec{I} & C_1 & \overleftarrow{x_i^1} \\ & C_2 & \overleftarrow{x_i^2} \\ & \vdots & \vdots \\ & C_n & \overleftarrow{x_i^n} \end{bmatrix}$$，则：$\overleftarrow{x_i^j} = \max_Q^j + \min_Q^j - x_i^j$。

在迭代过程中，小生境的边界 $x_i^j \in [\min_Q^j, \max_Q^j]$ 会根据小生境包含的群体规模动态调整，使搜索空间逐步缩小，从而提高收敛速度。若 $\overleftarrow{x_i^j} \notin [\min_Q^j, \max_Q^j]$，则说明反向粒子不在可行域内，本文对其随机重置，令 $\overleftarrow{x_i^j} = \psi \min_j + (1-\psi)\max_j$，其中 $\psi \in [0,1]$ 为随机数。

精英反向粒子的引入，有效拓展了搜索空间，在一定程度打破原小生境，促进种群进化。本文利用迭代次数确定精英反向粒子的引入时机，即每迭代 N 次，算法进行一次精英粒子反向学习，保证在一定概率范围内引导种群向更优位置进化。

4.5 算法流程

根据上述分析，基于小生境量子粒子群算法的特征选择方法主要步骤如下：

①设置种群规模 M，最大迭代次数 G，粒子最大速度 $V_{\max}$，共享距离 D 等变量参数；

②初始化种群——通过二进制编码产生初始种群，初始迭代次数 $g = 1$；

③选出 k 个样本作为中心，根据第3节的方法构造小生境；

④计算 $\overline{x(g)}$ 求出个体适应度值 $x_i(g)$，得出 $x_{i-best}(g)$、$x_{best}(k)$、$\tilde{x}(g)$；

⑤粒子位置更新，$x_i(g) = x_i(g+1)$，$g = g+1$；

⑥判断是否陷入局部最优，若是，对共享距离 D 内的粒子实施适应度动态共享；否则，转入步骤⑦；

⑦判断是否达到精英反向学习条件，若满足，则对最优粒子进行反向学习；否则，转入步骤⑧；

⑧判断是否满足迭代终止条件，若满足，将得到的最优个体输出，其对应的特征集即最终的求解结果；否则，转入步骤④。

5 实验与分析

5.1 实验设计

(1) 文本分类语料库

要通过实验对文本分类的效果进行验证，首先需要选取合适的语料库，其标准为使用广泛、标准规范、科学权威，这样便于实验数据的分析，确保实验结果与同行研究内容的可比性。本文参照文献[13]，将复旦大学中文文本分类语料库作为实验原始数据来源，相关数据可通过互联网获取，来源网址：http://www.nlpir.org/download/tc-corpus-answer.rar。该库中共有 20 个类别的文档，各类文档分布是非均匀的；文档又分

为训练集和测试集，训练语料共 9 804 篇，测试语料 9 833 篇，两者比例基本相当，剔除重复和损坏的文档后，训练集包含文档 8 214 篇，测试集包含文档 6 164 篇；每一个文档都有唯一的文件名。本实验从语料库的 20 个类别中选择出 10 类，各类文档数如表 1 所示。

表 1　语料库文档数目

类别	训练文档数	测试文档数
计算机	628	591
太空	506	248
军事	74	75
体育	584	489
历史	466	468
政治	573	482
经济	480	419
艺术	510	286
农业	547	435
环境	405	371

(2) 实验环境及参数

实验在 Intel(R) Core(TM) i7-4790 CPU@3.60GHz 计算机平台上，利用 64 位 Windows 7 操作系统实现，通过 ICTCLAS 系统进行分词处理，使用 Java 语言开发的 Weka 软件完成数据处理操作。Weka 软件是数据分析常用的软件之一，其中包含数据处理、回归分析、聚类与分类、可视化等不同的功能模块和工具，使用较为方便。

实验参数设置：粒子规模 $M=60$，最大迭代次数 $G=1000$。为评估算法性能，实验先后采用本文所提出的 NOL-QPSO 和 IG、MI 及 QPSO 等 4 种不同的特征选择方法对语料库中的文档进行特征选择，待特征选择后，使用 KNN 分类器对文本进行分类，设定分类器参数 $K=10$。为对分类结果定量分析，通过准确率 P、召回率 R 和综合评价指标 F1 值对各种方法的分类效果进行对比，判别结果说明如表 2 所示，计算方法如下：

$$P=a/(a+b)$$

$$R=a/(a+c)$$

$$F1=2pr/(p+r)$$

表 2　判别结果说明

类别	判断属此类	判断不属此类
判断属此类	a	b
判断不属此类	c	d

5.2　实验结果及分析

实验结果如表 3 所示，数据分别反映出通过 4 种方法对文本分类所得结果的准确率、召回率和 F1 值。

表 3　实验结果

类别	NOL-QPSO			IG			MI			QPSO		
	P(%)	R(%)	F1 值(%)	P(%)	R(%)	F1 值(%)	P(%)	R(%)	F1 值	P(%)	R(%)	F1 值(%)
计算机	94.26	93.88	94.07	85.24	82.46	83.83	81.52	85.49	83.46	80.04	76.52	78.24
太空	95.21	94.54	94.87	80.59	78.96	79.77	80.92	82.57	81.74	75.83	77.20	76.51
军事	94..27	93.56	93.91	76.42	80.12	78.23	83.10	79.86	81.45	76.44	72.56	74.45
体育	93.58	94.08	93.83	84.46	85.60	85.03	79.56	81.54	80.54	69.38	76.17	72.62
历史	92.25	93.50	92.87	82.42	81.86	82.14	82.06	80.46	81.25	72.56	71.39	71.97
政治	90.10	91.92	91.00	80.88	82.43	81.65	74.28	78.54	76.35	75.18	78.66	76.88
经济	94.73	93.52	94.12	84.26	80.85	82.52	81.72	85.22	83.43	76.29	72.36	74.27
艺术	94.20	90.84	92.49	88.24	84.96	86.57	82.91	78.53	80.66	76.80	71.22	73.90
农业	95.78	94.22	94.99	80.56	76.84	78.66	80.48	79.31	79.89	67.12	76.18	71.36
环境	92.46	90.68	91.56	76.85	80.47	78.62	78.19	67.12	72.23	81.03	80.56	80.79
均值	93.684	93.074	93.378	81.992	81.455	81.723	80.474	79.864	80.168	75.067	75.282	75.174

图 1 是根据表 3 数据得到的条形图，更直观地反映了不同特征选择方法之间分类效果的差别。不难看出，通过不同方法对文本进行分类，所得结果的准确率和召回率不同，本文提出的基于 NOL-QPSO 的特征选择方法综合考虑了特征之间的相互关系，在文本分类中表现出更好的精确性；与 QPSO 算法相比，

NOL-QPSO 改善了算法易陷入局部最优的缺点，寻优精度更高；由于本实验所选语料库中的样本分布是不均匀的，所以 IG 方法的分类效果受到一定影响。图 2 则分别对 4 种不同方法在文本分类中各类别的准确率和召回率依次作出比较，从结果可以直观地看出，NOL-QPSO 算法与其他三种方法相比，不仅准确率和召回率更高，而且性能更为稳定。根据实验结果对 4 种方法按从优到劣依次排序为：NOL-QPSO > MI > IG > QPSO。

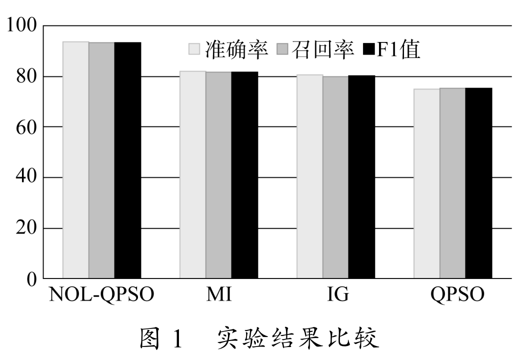
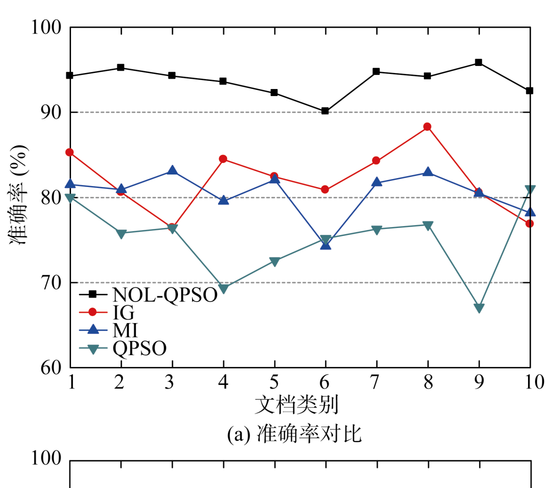
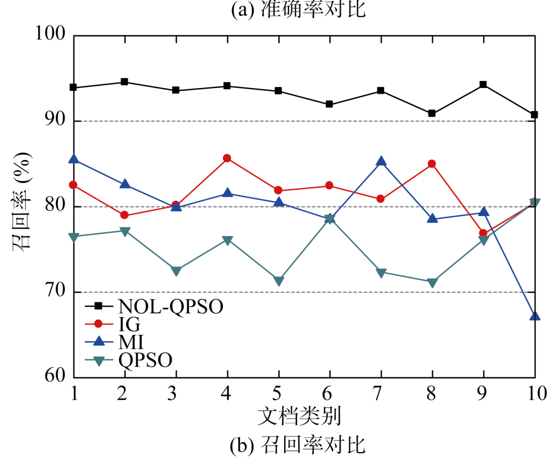


图 1　实验结果比较



(a) 准确率对比



(b) 召回率对比

图 2　4 种方法性能对比

表 4 反映了各算法的平均运行时间，可以看出，本文提出的算法与其他算法在时间复杂度上差别不大，在时间允许范围之内取得了更好的分类效果，也证明了 NOL-QPSO 方法的有效性。

表 4　算法平均运行时间

所用方法	NOL-QPSO	MI	IG	QPSO
运行时间(s)	1 744	1 541	1 496	1 598

6　结　语

本文以可拓理论为基础构造算法模型，结合适应度共享和精英反向学习等策略，提出一种用于特征选择的改进粒子群算法 NOL-QPSO，主要工作体现在以下几点：

(1) 以可拓理论为基础，将小生境策略用于量子粒子群算法，以改善算法的全局性；

(2) 加入适应度动态共享环节，通过引入共享函数调整适应度，避免陷入早熟；

(3) 通过实验，对不同的文本特征选择方法进行比较，验证了本文所提算法的有效性。

本文方法在文本分类中取得了较好的分类效果，而时间效率与传统方法基本相当，与 QPSO 算法相比，算法搜索精度得到明显提升，为中文文本的特征选择问题提供了一种方法和思路。对算法时间效率的改善，是后续的研究重点之一。

参考文献：

[1] 何熊熊，管俊轶，叶宣佐. 一种基于密度和网格的簇心可确定聚类算法[J]. 控制与决策，2017，32(5)：913-919.（He Xiongxiong, Guan Junyi, Ye Xuanzuo. A Density-based and Grid-based Cluster Centers Determination Clustering Algorithm[J]. Control and Decision, 2017, 32(5): 913-919.）

[2] 任俊亮，邢清华，李强，等. 采用自适应概率粒子群算法的反导预警资源调度方法[J]. 空军工程大学学报：自然科学版，2014，15(6)：45-48.（Ren Junliang, Xing Qinghua, Li Qiang, et al. Resource Scheduling Method of Missile Defense Early Warning System Based on Self-Adaptive Probability Particle Swam Optimization [J]. Journal of Air Force Engineering University: Natural Science Edition, 2014, 15(6): 45-48.）

[3] Sun J, Feng B, Xu W B. Particle Swarm Optimization with Particle Having Quantum Behavior [C]//Proceedings of

Congress on Evolutionary Computation, Portland.USA: IEEE Press, 2004, 1: 325-331.

[4] Sun J, Xu W B, Feng B. Adaptive Parameter Control for Quantum Behaved Particle Swarm Optimization on Individual Level[C]//Proceedings of IEEE International Conference on Systems, Man and Cybernetics. Piscataway: IEEE Press, 2005: 3049-3054.

[5] 路永和, 梁明辉. 遗传算法在改进文本特征提取方法中的应用[J]. 现代图书情报技术, 2014(4): 48-57. (Lu Yonghe, Liang Minghui. Improvement of Text Feature Extraction with Genetic Algorithm [J]. New Technology of Library and Information Service, 2014(4): 48-57.)

[6] 刘逵, 周竹荣. 基于野草算法的文本特征选择[J]. 计算机应用, 2012, 32(8): 2245-2249. (Liu Kui, Zhou Zhurong. Text Feature Selection Method Based on Invasive Weed Optimization [J]. Journal of Computer Applications, 2012, 32(8): 2245-2249.)

[7] 林令娟, 刘希玉. 基于微粒群优化的快速 K-近邻分类算法[J]. 山东科学, 2009, 22(1): 13-16. (Lin Lingjuan, Liu Xiyu. A Particle Swarm Optimization Based Rapid K-nearest Neighbor Classification Algorithm [J]. Shandong Science, 2009, 22(1): 13-16.)

[8] 李欢, 焦建民. 简化的粒子群优化快速 KNN 分类算法[J]. 计算机工程与应用, 2008, 44(32): 57-59. (Li Huan, Jiao Jianmin. Improved Simplified PSO KNN Classification Algorithm[J]. Computer Engineering and Applications, 2008, 44(32): 57-59.)

[9] 拓守恒. 基于改进 PSO 的 SVM 文本分类研究[J]. 电脑开发与应用, 2010, 23(10): 3-5, 8. (Tuo Shouheng. Research on Text Categorization Based on Support Vector Machine Optimized by Particle Swarm Optimization Algorithm[J]. Computer Development & Applications, 2010, 23(10): 3-5, 8.)

[10] 孙洋. 粒子群算法的改进及其在文本分类上的应用[J]. 中央民族大学学报: 自然科学版, 2008, 17(3): 57-62. (Sun Yang. The Improvement of PSO Algorithm and Application of Text Classifier[J]. Journal of the Central University for Nationalities: Natural Sciences Edition, 2008, 17(3): 57-62.)

[11] 徐辉. 基于混沌二进制粒子群优化的 KNN 文本分类算法[J]. 微电子学与计算机, 2012, 29(8): 204-208. (Xu Hui. KNN Text Classification Algorithm Based on Chaotic Binary Particle Swarm Optimization [J]. Microelectronics & Computer, 2012, 29(8): 204-208.)

[12] 谭德坤. 基于混沌微粒群算法的文本分类研究[J]. 计算机应用研究, 2010, 27(12): 4464-4466. (Tan Dekun. Research of Chinese Text Categorization Based on Chaotic Particle Swarm Optimization[J]. Application Research of Computers, 2010, 27(12): 4464-4466.)

[13] 朱颢东, 钟勇. 基于并行二进制免疫量子粒子群优化的特征选择方法[J]. 控制与决策, 2010, 25(1): 53-63. (Zhu Haodong, Zhong Yong. Feature Selection Method Based on PBIQPSO[J]. Control and Decision, 2010, 25(1): 53-63.)

[14] 孔莉芳, 张虹. 用于特征子集选择的异步并行微粒群优化方法[J]. 控制与决策, 2012, 27(7): 967-973. (Kong Lifang, Zhang Hong. Asynchronous Parallel Particle Swarm Optimizer for Feature Subset Selection [J]. Control and Decision, 2012, 27(7): 967-973.)

[15] 巩敦卫, 胡滢, 张勇. 基于多目标微粒群优化的异质数据特征选择[J]. 电子学报, 2014, 42(7): 1320-1326. (Gong Dunwei, Hu Ying, Zhang Yong. Feature Selection of Heterogeneous Data Based on Multi-objective Particle Swarm Optimization[J]. Acta Electronica Sinica, 2014, 42(7): 1320-1326.)

[16] 付强, 王刚, 王明宇, 等. 基于小生境遗传算法的制导雷达误差估计[J]. 空军工程大学学报: 自然科学版, 2011, 11(6): 50-53. (Fu Qiang, Wang Gang, Wang Mingyu, et al. Research of Guidance Radar Error Estimation Based on the Niche Genetic Algorithm[J]. Journal of Air Force Engineering University: Natural Science Edition, 2011, 11(6): 50-53.)

[17] 杨春燕, 蔡文. 可拓学[M]. 北京: 科学出版社, 2014: 18-96. (Yang Chunyan, Cai Wen. Extenics[M]. Beijing: Science Press, 2014: 18-96.)

[18] 赵敏, 林道荣, 瞿波, 等. 一种新的基于小生境模拟退火的遗传算法[J].辽宁工程技术大学学报: 自然科学版, 2013, 32(3): 367-372. (Zhao Min, Lin Daorong, Qu Bo, et al. A New Genetic Algorithm Based on Niche Simulated Annealing [J]. Journal of Liaoning Technical University: Natural Science, 2013, 32(3): 367-372.)

[19] 李中华, 张泰山. 可拓聚类适应度共享小生境遗传算法研究[J]. 哈尔滨工业大学学报, 2016, 48(5): 178-183. (Li Zhonghua, Zhang Taishan. Research of Fitness Sharing Niche Genetic Algorithms Based on Extension Clustering [J]. Journal of Harbin Institute of Technology, 2016, 48(5): 178-183.)

[20] 曾维宏. 基于粗糙集理论的数据挖掘算法研究[D]. 郑州: 郑州大学, 2005. (Zeng Weihong. Research of Reduction Algorithm Based on Rough Set Theory [D]. Zhengzhou: Zhengzhou University, 2005.)

[21] 张珂, 黄永峰, 李星. 一种基于适应度和节点聚类的 P2P

拓扑建模方法[J]. 电子学报, 2010, 38(7): 1634-1640. (Zhang Ke, Huang Yongfeng, Li Xing. A Model for Topology of P2P Network Based on Fitness and Node Clustering[J]. Acta Electronica Sinica, 2010, 38(7): 1634-1640.)

[22] 谭熠峰, 孙婷婷, 徐新民. 基于动态因子和共享适应度的改进粒子群算法[J]. 浙江大学学报: 理学版, 2016, 43(6): 696-700. (Tan Yifeng, Sun Tingting, Xu Xinmin. A Modified Particle Swarm Optimization Algorithm Based on Dynamic Learning Factors and Sharing Method[J]. Journal of Zhejiang University: Science Edition, 2016, 43(6): 696-700.)

[23] 邵鹏, 吴志健, 周炫余, 等. 基于折射原理反向学习模型的改进粒子群算法[J]. 电子学报, 2015, 43(11): 2137-2144. (Shao Peng, Wu Zhijian, Zhou Xuanyu, et al. Improved Particle Swarm Optimization Algorithm Based on Opposite Learning of Refraction[J]. Acta Electronica Sinica, 2015, 43(11): 2137-2144.)

作者贡献声明:

李志鹏: 算法设计, 论文起草、修改, 数据收集, 实验分析;
李卫忠: 确定研究方向, 提供算法设计方案, 提出论文修改意见。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: lizhipeng0888@yeah.net。
[1] 李志鹏. 语料库.zip. 实验语料.
[2] 李志鹏. 算法代码.zip. 算法实现.
[3] 李志鹏. 实验数据.xls. 实验过程数据.

收稿日期: 2017-05-27
收修改稿日期: 2017-07-10

Feature Selection Based on Modified QPSO Algorithm

Li Zhipeng Li Weizhong
(Air and Missile Defense College, Air Force Engineering University, Xi’an 710051, China)

Abstract: [**Objective**] This study proposes an algorithm for feature selection aiming to improve the precision and efficiency of text classification. [**Methods**] First, we selected features based on their characteristics. Then, we constructed the algorithm with extension theory to strengthen its searching ability. Finally, we compared the performance of different methods for text classification. [**Results**] Compared with IG, MI and QPSO, the proposed algorithm had better accuracy in feature selection. [**Limitations**] The efficiency of our algorithm needs to be improved. [**Conclusions**] The modified QPSO Algorithm is an effective way to select features.

Keywords: Feature Selection Quantum-behaved Particle Swarm Extenics Niche Fitness Sharing